

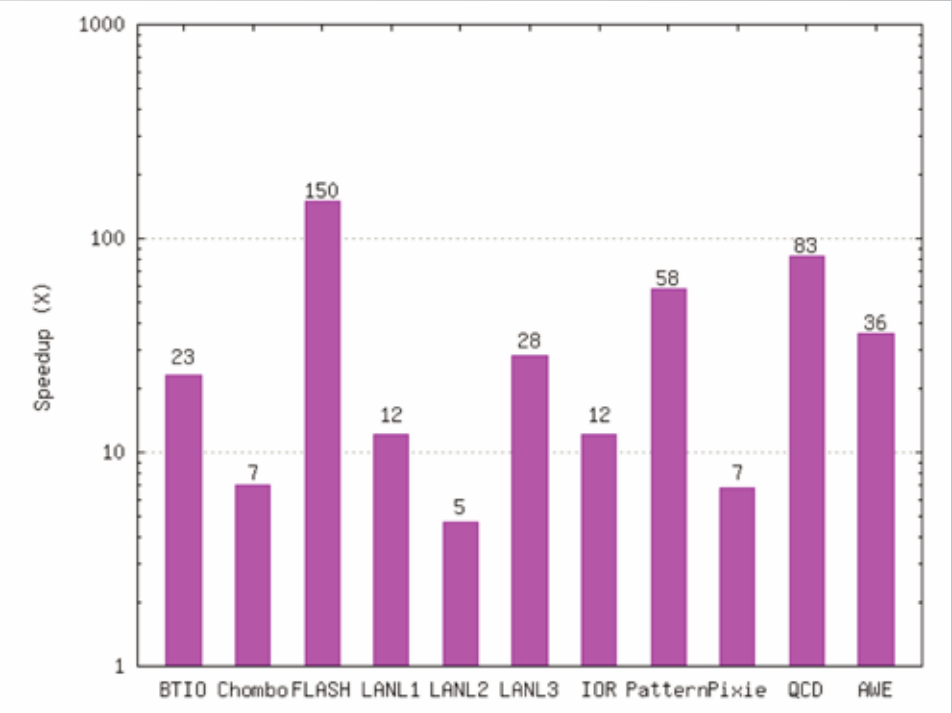
PLFS Updates and Ongoing Research

What: PLFS: The Parallel Log-structured File System

- High bandwidth parallel write IO regardless of workload
- Virtual file system
- Transparent, low overhead, portable
- POSIX interface for unmodified applications
- MPI-IO interface for applications using MPI-IO
- PLFS API for applications that want more low-level control

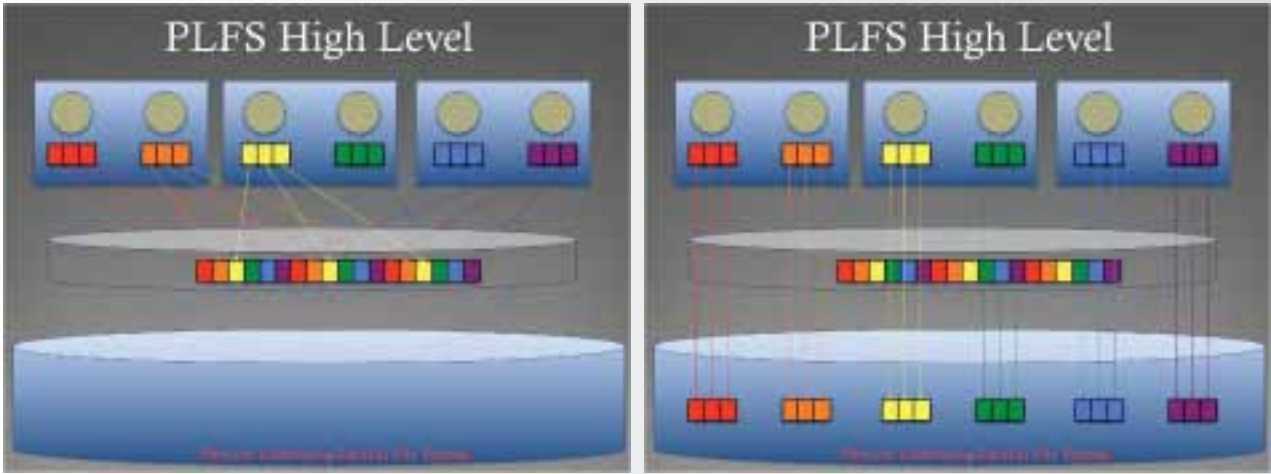
Why

- Applications writing to a shared file (N-1) perform very poorly in Lustre, GPFS, and PanFS
- Applications writing to unique files (N-N) perform very well in Lustre, GPFS, and PanFS
- PLFS transparently converts N-1 into N-N
- No modifications required to existing applications. ls, tar, grep, etc all just work as expected



How: Low-level architecture and implementation details

- On a hidden physical storage system, PLFS creates a directory for each logical file
- Each writing process writes data to a log file and metadata to an index file
- On reads and stats, the index files are read to reconstruct the logical file

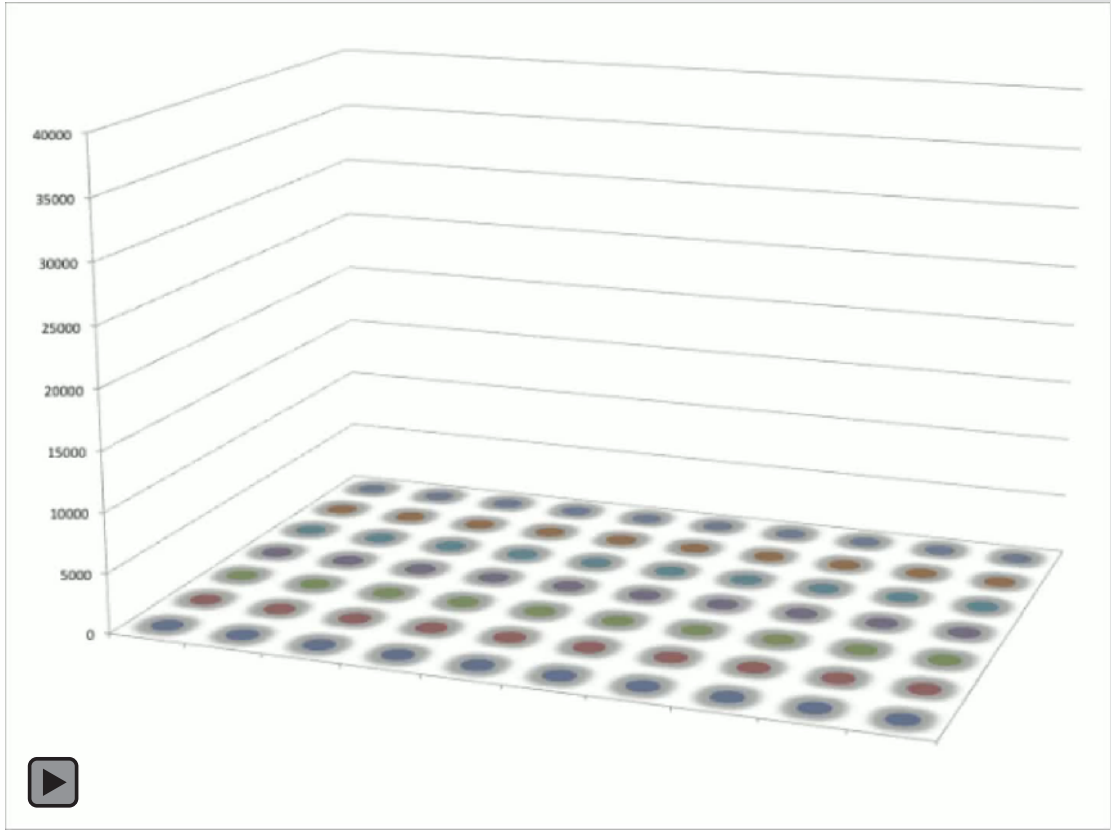


Status

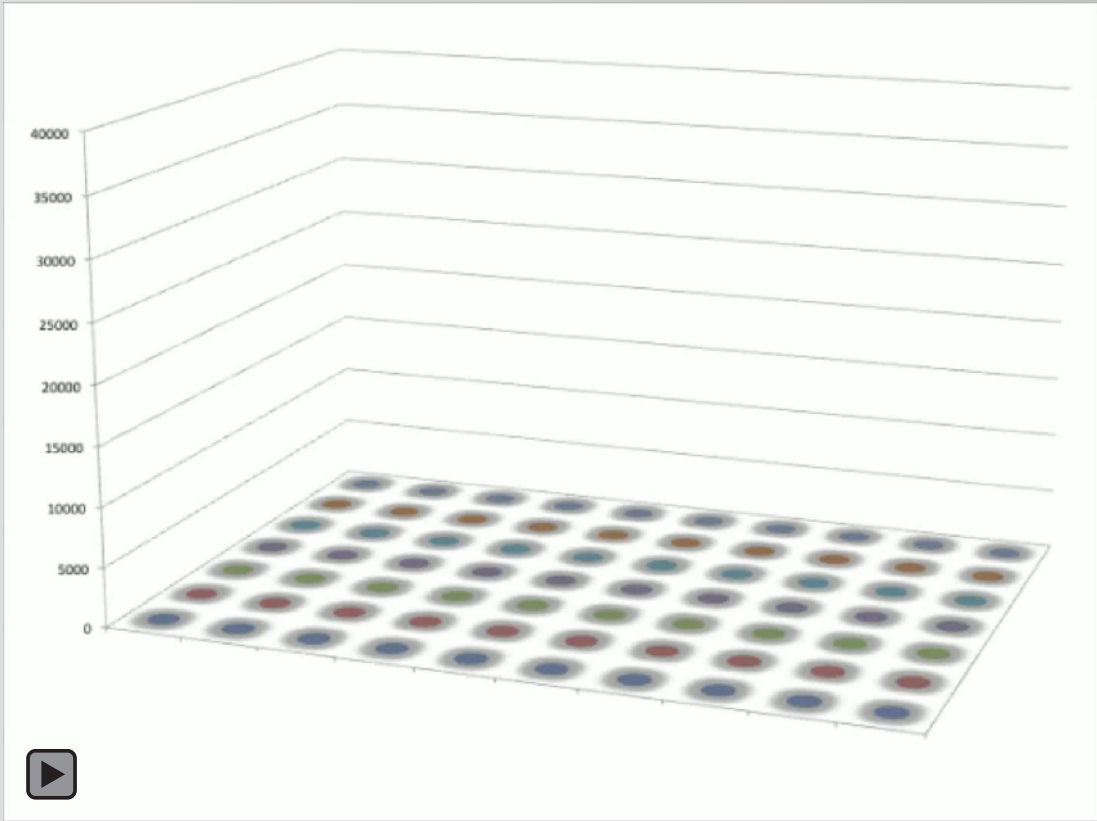
- POSIX and MPI-IO interfaces installed on LANL Roadrunner
- MPI-IO interface being evaluated with good results on ORNL Jaguar
- POSIX and MPI-IO interfaces in early evaluation at the United Kingdom’s AWE
- Available as part of Tri-lab software stack (TOSS)
- Source available: <http://sourceforge.net/projects/plfs>

Current Efforts

- As expected for a log-structured file system, some read workloads present challenges
- Exploring read optimizations in the MPI-IO interface and the POSIX as well
- Consideration of data integrity checks for almost end-to-end reliability
- Compression to speed metadata operations and reduce storage usage
- Thread pools for increasing concurrency when storage and CPU are underutilized



Disk activity without PLFS





Disk activity with PLFS

LA-UR 10-07443

John Bent, johnbent@lanl.gov, 505-663-5820

PRObE: Parallel Reconfigurable Observational Environment





An NSF-sponsored computer systems research center

•The PRObE project will provide two classes of computing resources: large research clusters and unique and advanced hardware.

•The large clusters will be repurposed machines from DOE facilities. The machines will be older generation technology researchers can use to do large-scale research.

•Researchers will have dedicated use of a complete cluster to execute research.

A highly reconfigurable, remotely accessible and controllable environment dedicated to systems research. We envision this unique system will support research in many systems related fields such as Operating Systems, Storage, and High End Computing.

Researchers will have complete control of the hardware while they are running experiments.

Researchers can inject both hardware and software failures while monitoring the system to see how it reacts to such failures.

PRObE, at full production scale, will provide at least two 1024 node clusters, one of 200 nodes, and some smaller machines with extreme core count and bleeding edge technology.

The PRObE research environment will be based on Emulab testbed-management software.





<http://newmexicoconsortium.org/probe>
 Email: probe@newmexicoconsortium.org

Information Science and Technology Institute

Increasing Our Technical Competitiveness Through University Collaborations



A family of strategic partnerships and collaborations with leading research universities and government institutions and organizations. ISTI was formed to support IS&T at LANL broadly to meet the decadal challenges in the information, computer, computational, and knowledge sciences.



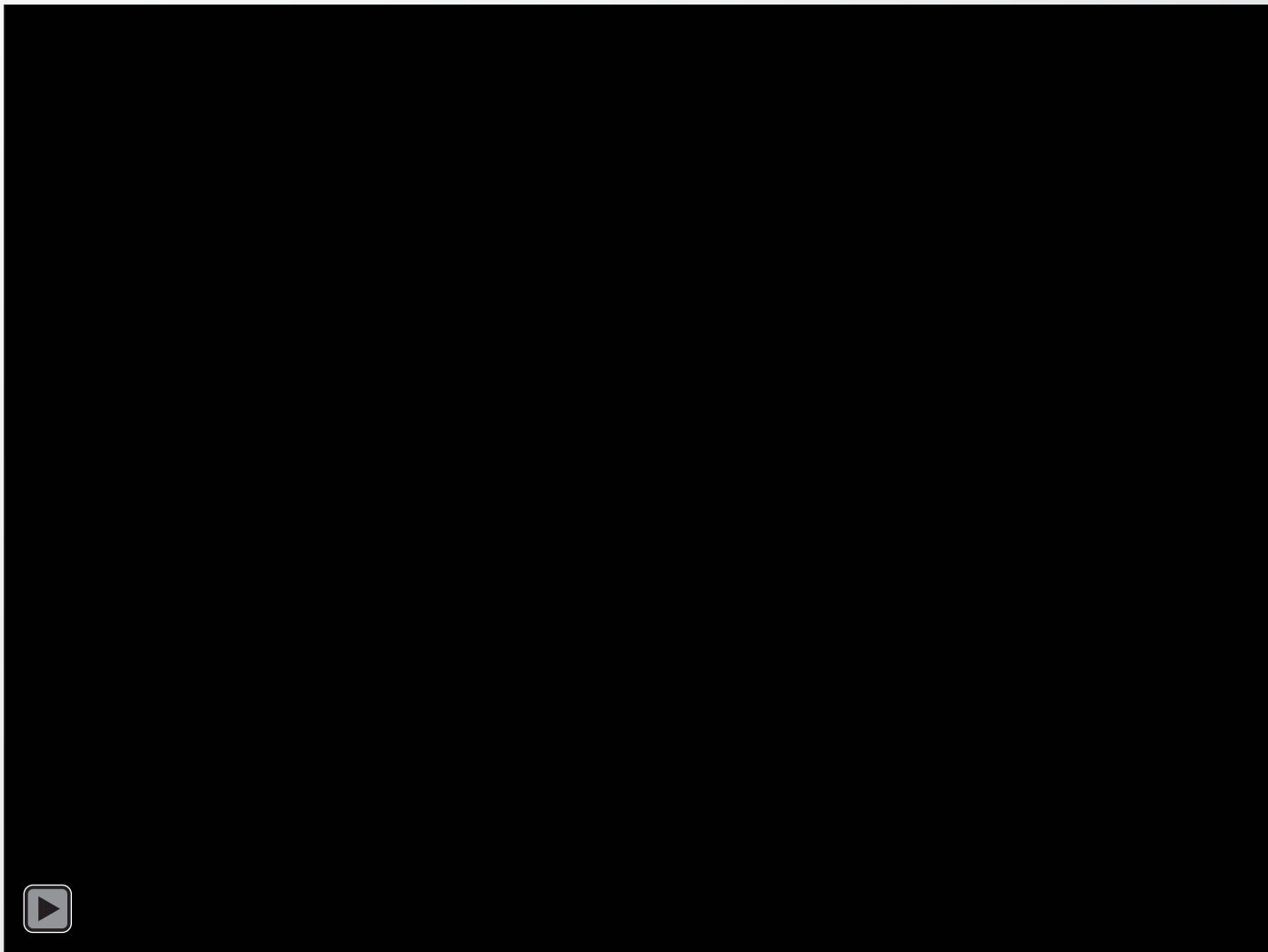
ISTI Goals

- Meet the next decade’s top challenges in the information, computer, computational, and knowledge sciences.
- Partner with leading research universities in targeted challenge areas
- Foster strong technical collaboration and collaborative research
- Work with Directorates and line orgs to support programmatic goals and to develop capability in strategic science & technology areas.
- Provide networking, research, revitalization, and educational opportunities to LANL staff.
- Provide leadership outreach from LANL to the national community



Computer System, Cluster, & Networking Summer Institute

An NSF-funded undergraduate summer intensive



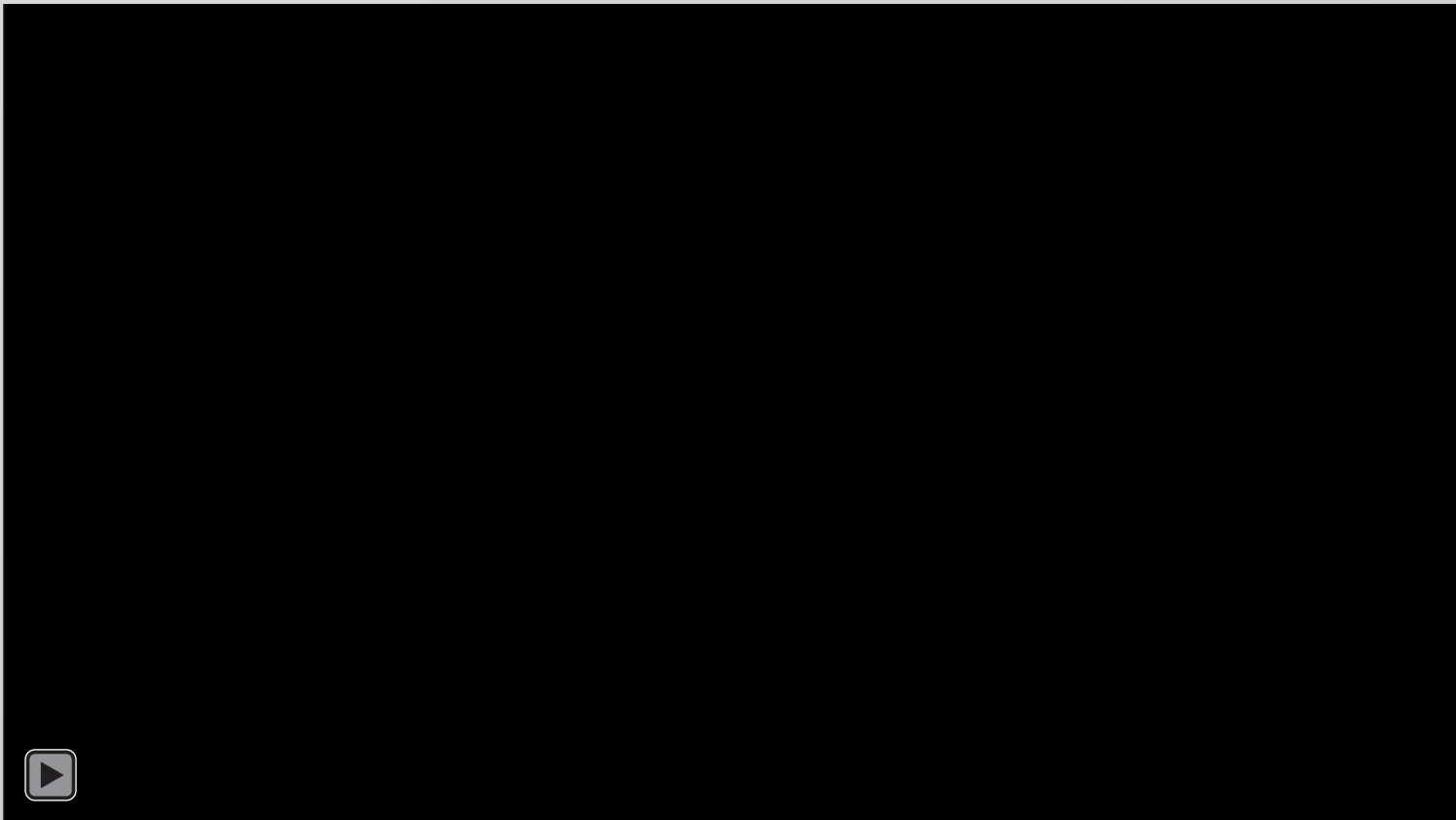
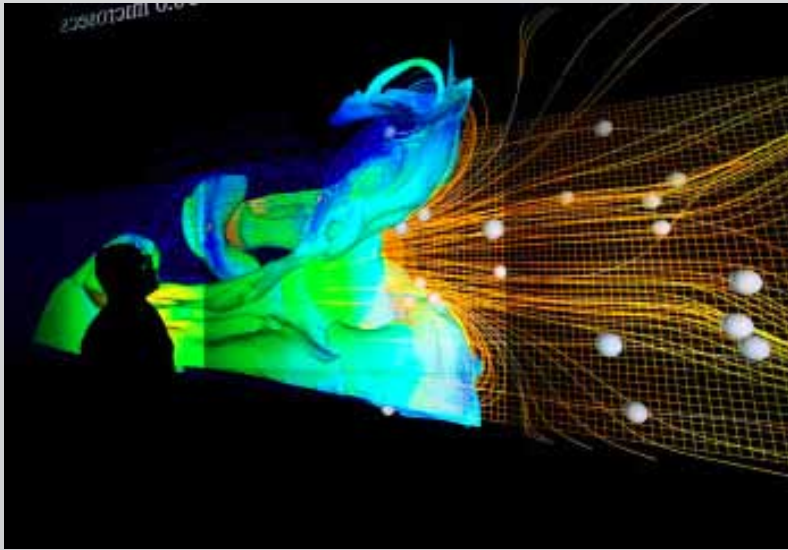
Highly-selective, 9-week program designed for third year (i.e., Junior) undergraduates, although Sophomores and Seniors may be considered.

Students work in small project teams to execute real-world projects on computer clusters that they have assembled and configured.

A university instructor provides class instruction

Subject matter experts from Los Alamos National Laboratory mentor the team projects which are presented at a technical symposium at the conclusion of the Summer Institute

Brochure and application details are available at:
institute.lanl.gov/isti/summer-school/cluster-network/

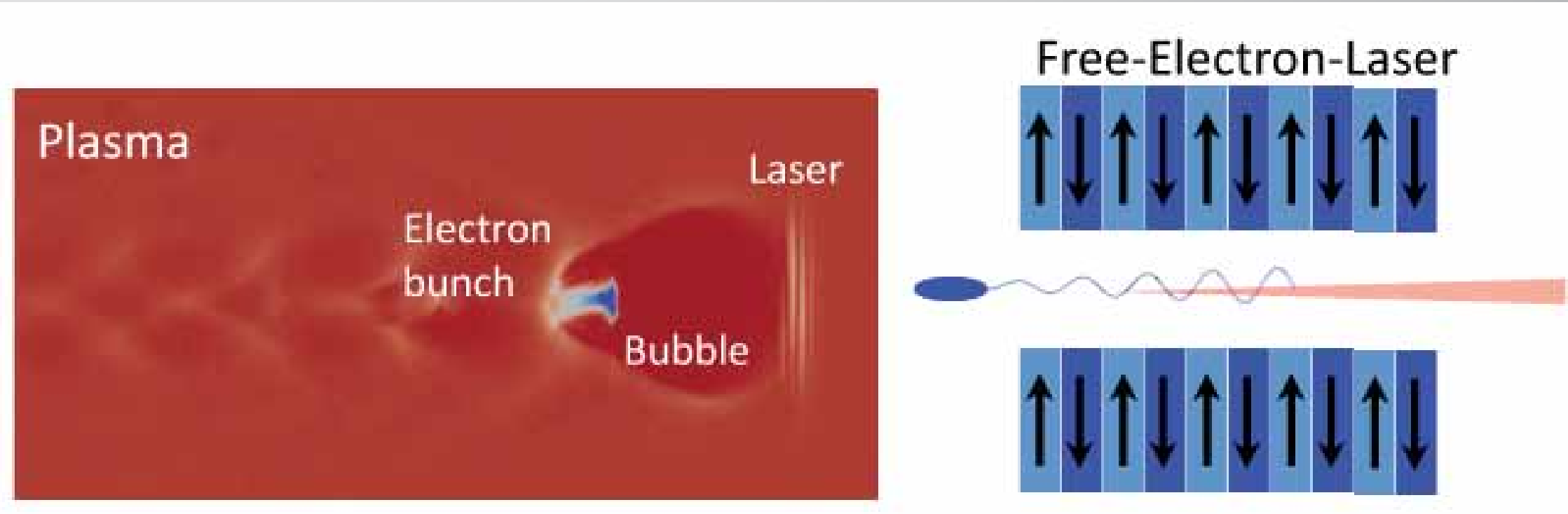


Carol Hogsett, carol@lanl.gov

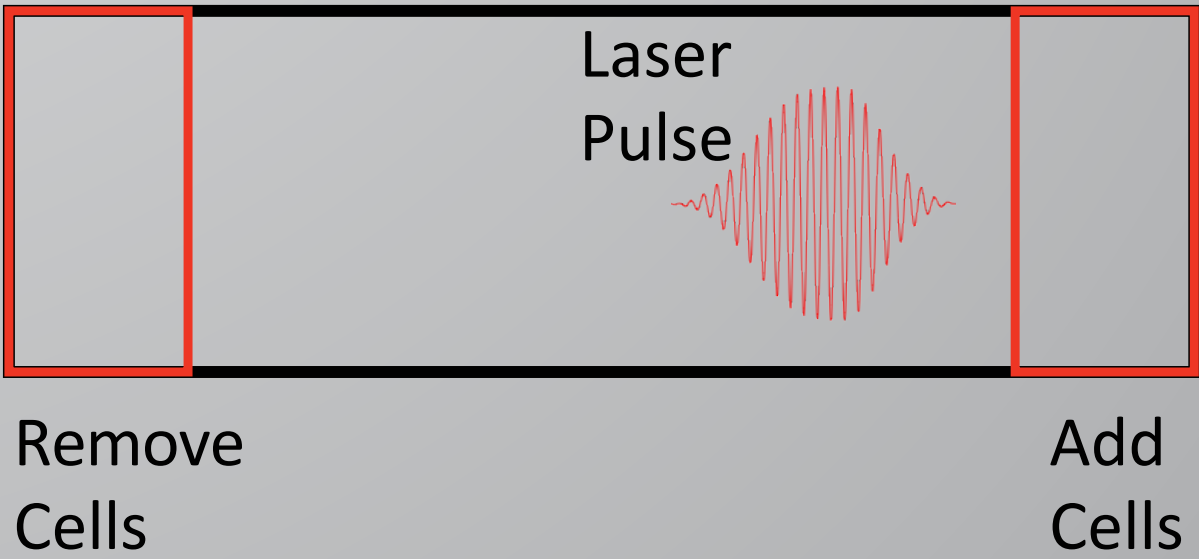
Applications will be accepted:
JAN. 10 — FEB. 18, 2011

Laser Driven Wakefield Accelerators

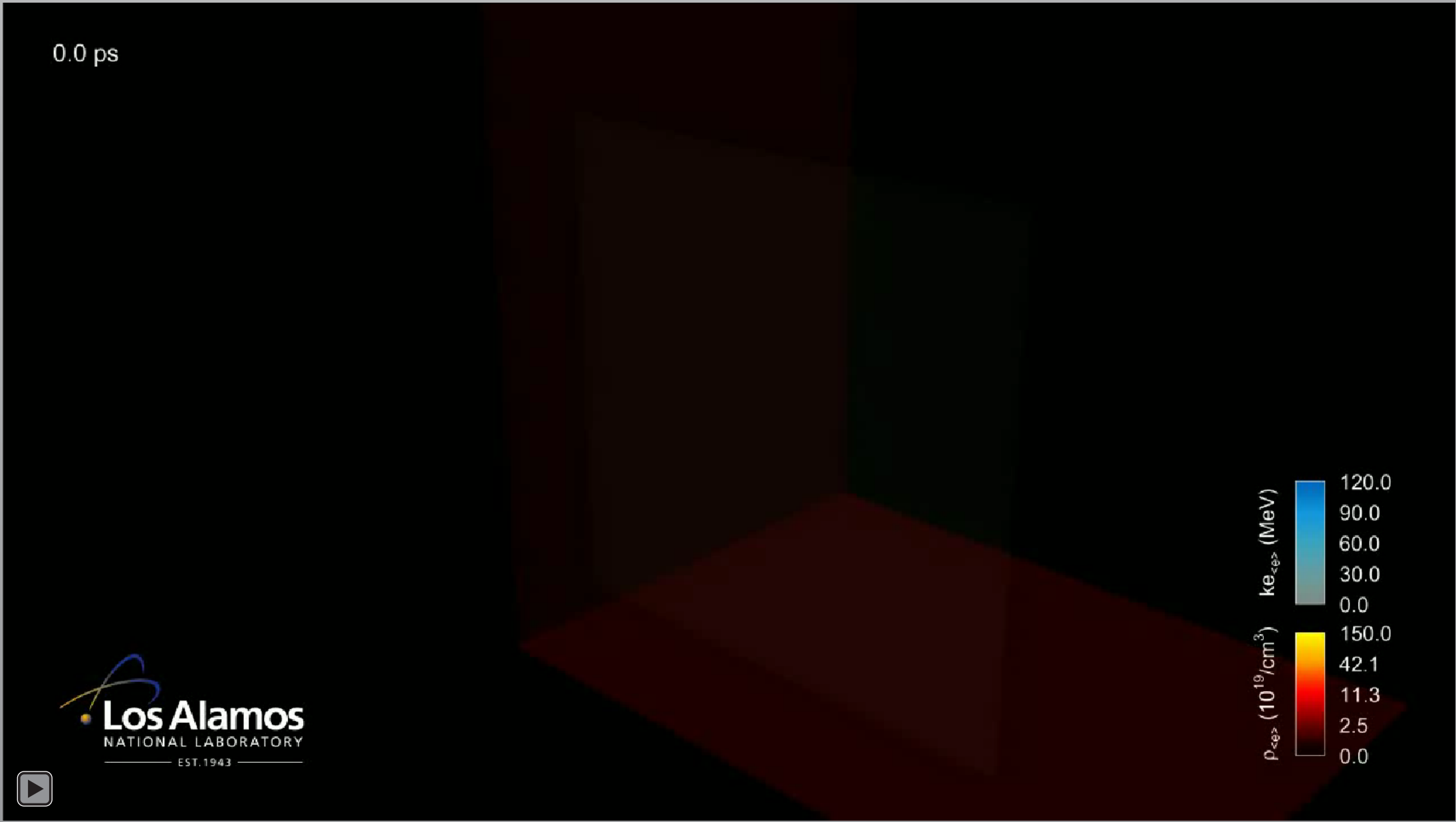
An intense laser pulse propagating though a plasma can blowout most of the electrons to form a bubble within the wakefield. Electrons trapped inside the bubble are accelerated to high energy and are attractive for driving an x-ray free-electron-laser (FEL) which may offer a powerful new diagnostic for imaging of nano-scale phenomena in materials & biology.



Domain moves to follow laser pulse & bubble



Example simulation illustrating moving window:



HPC Data Management

File Transfer Agents (FTAs)

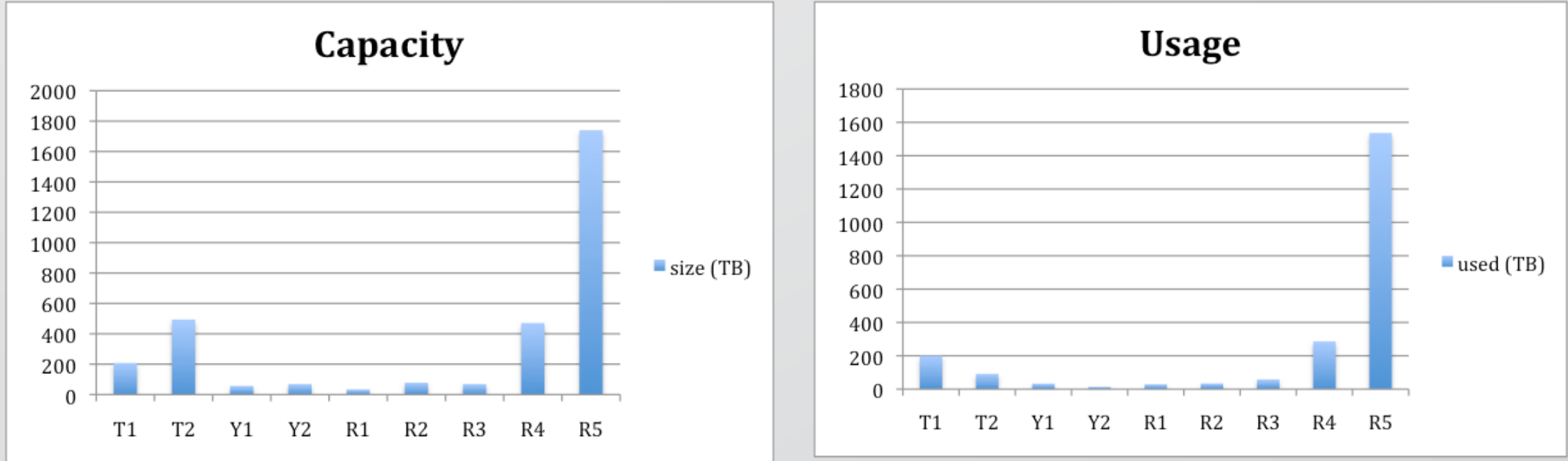
Parallel file systems work best with a distributed workload. LANL has special purpose clusters optimized for the distributed movement of parallel data sets from:

- one parallel file system to another
- parallel file system to archival storage
- archival storage to parallel file system
- between Tri-Lab sites.

Using a MOAB configuration, users submit jobs from any cluster to FTAs for data pre-staging before a job runs, then archival after the job finishes. Data set sizes are ever in-creasing and require distributed resources to move in a reasonable time.

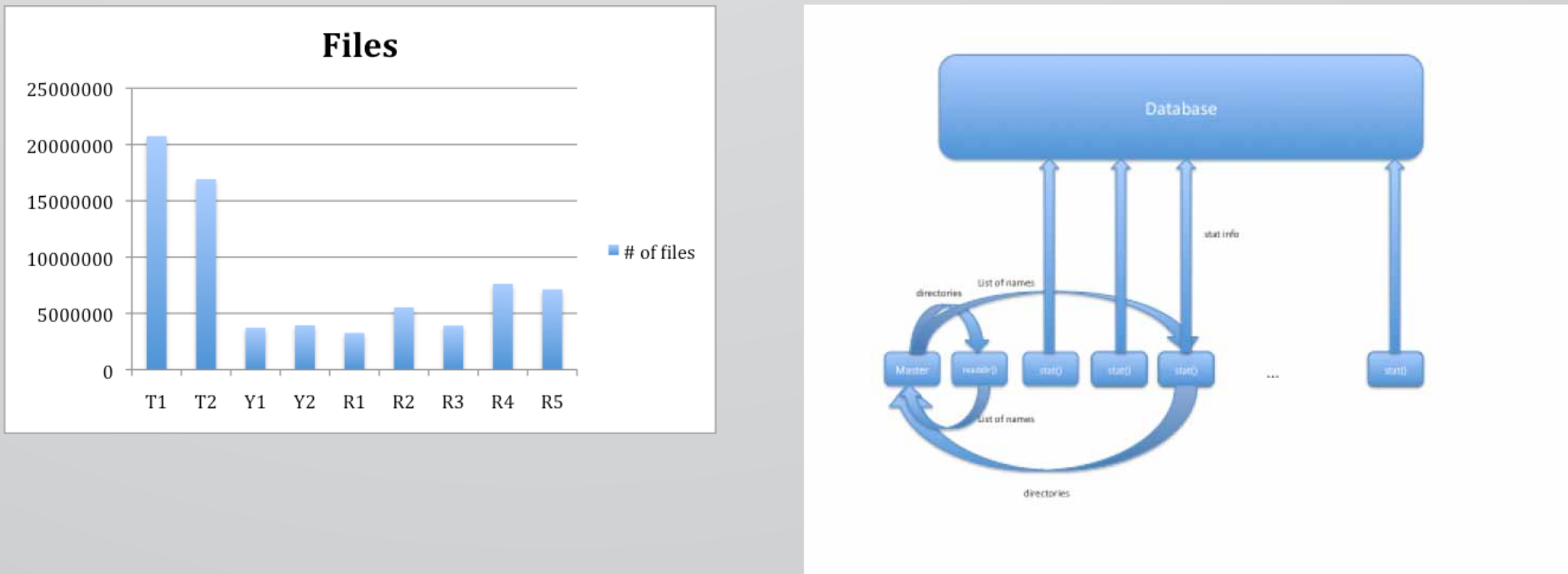
The parallel file system’s primary function is to capture checkpoint data to protect the large job from having to restart in the event of system failure. Clusters have an expected time for a job size before a job is interrupted due to failure. Because forward progress in a simulation is required, the time required for a checkpoint can only be a small percent-age of the expected runtime. Parallel file systems have to offer performance as well as available space. FTAs give the ability to move data to archival storage, allowing space for future checkpoints on the parallel file system.

Graphs show the increase over time from older to newer file systems for the Turquoise, Yellow, and Red network partitions.



Data Management Services (DMS)

The amount of data on the parallel file systems is reaching into the petabytes. These file systems regularly contain millions of files and directories. The I/O mode of user jobs varies. Graph shows a comparison of total files between the different file systems.



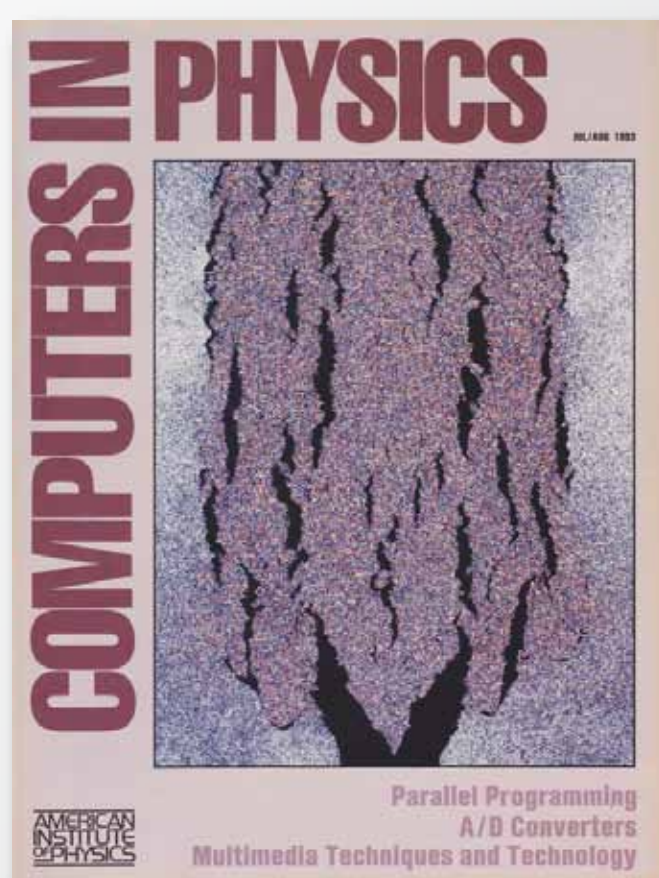
The database inserts files from the snapshot view into an expired files view if the file is old enough. Notifications are sent to users weekly for all files in the expired files view. A week later, the purger verifies if the file has not been updated on the file system and if not, removes the files. The purge process is a distributed job that runs on the DMS cluster to handle the large workload of files.

For more information, join our Google group at: <http://groups.google.com/group/hpc-monitoring>
 With code access at:
<https://sites.google.com/site/hpcmonitoring/best-practices/purger-lanl>

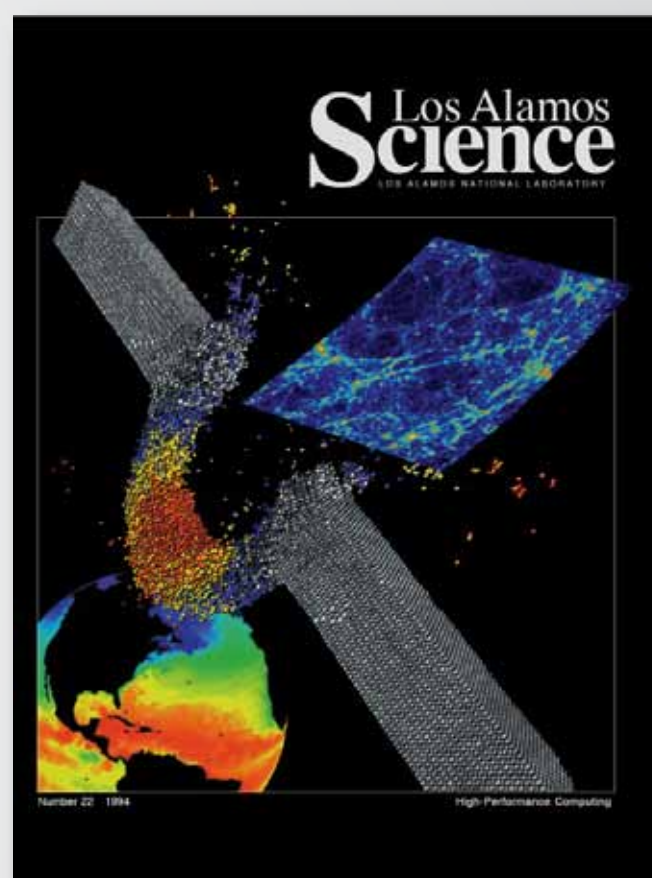
Ben McClelland (ben@lanl.gov)
 Gary Grider (ggrider@lanl.gov)
 Alfred Torrez (atorrez@lanl.gov)
 Aaron Torres (agtorre@lanl.gov)
 Dave Montoya (dmont@lanl.gov)

SPaSM*-Generated Cover Art

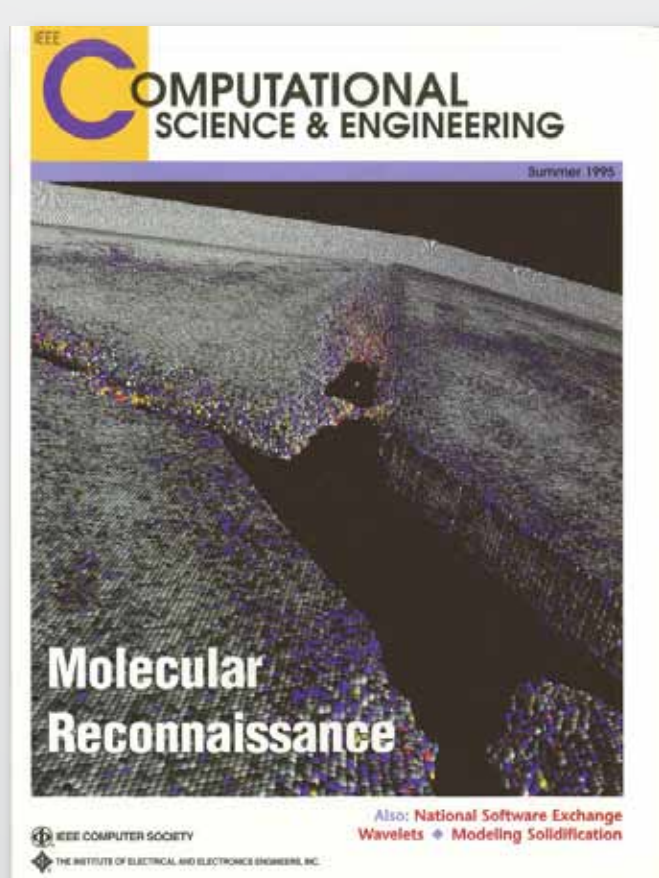
“*Scalable Parallel Short-range Molecular dynamics”



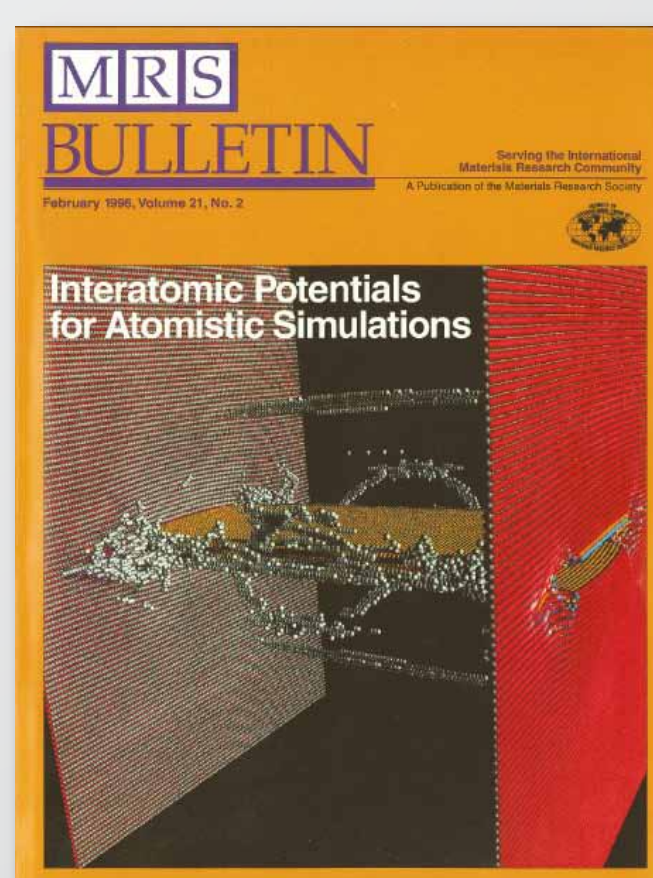
1993



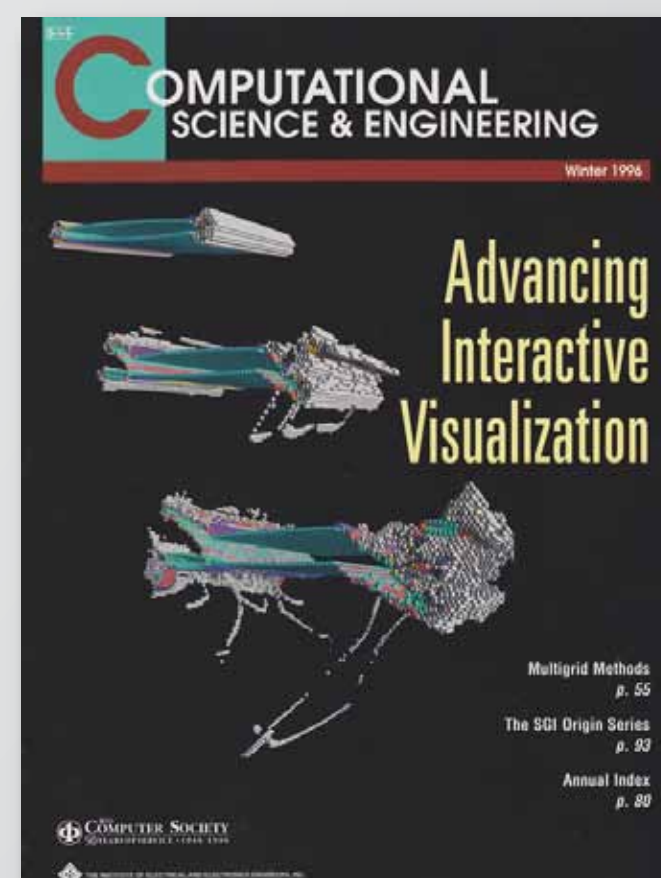
1994



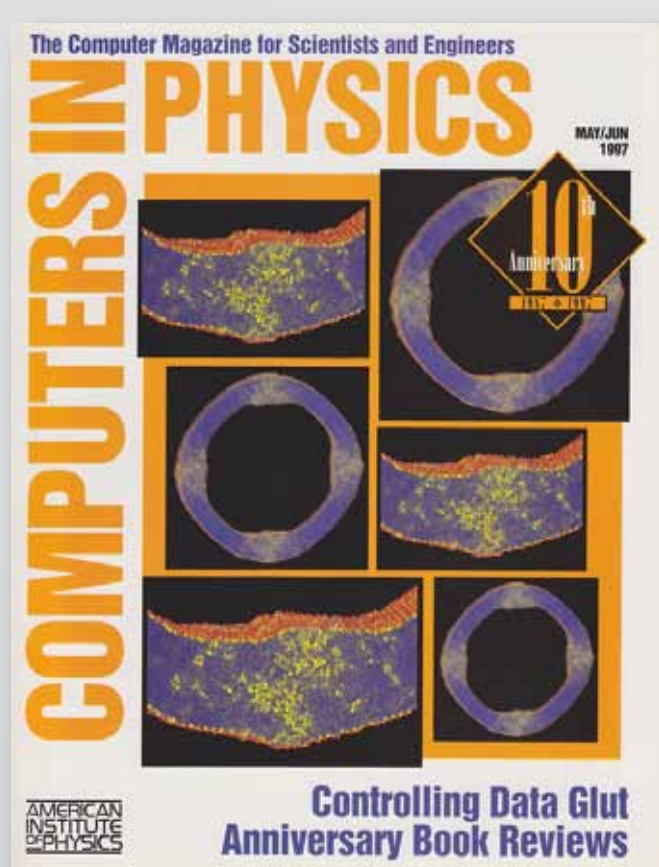
1995



1996



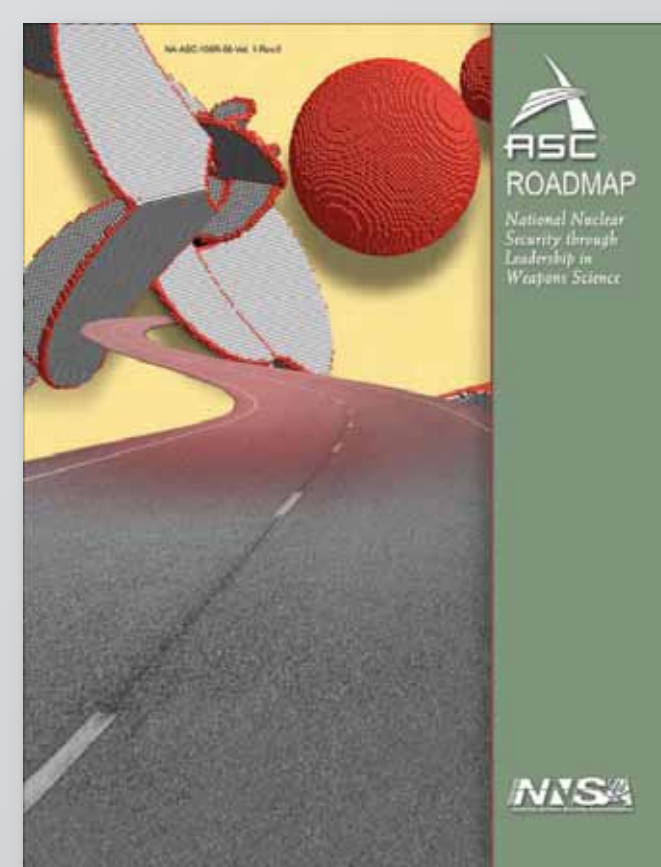
1996



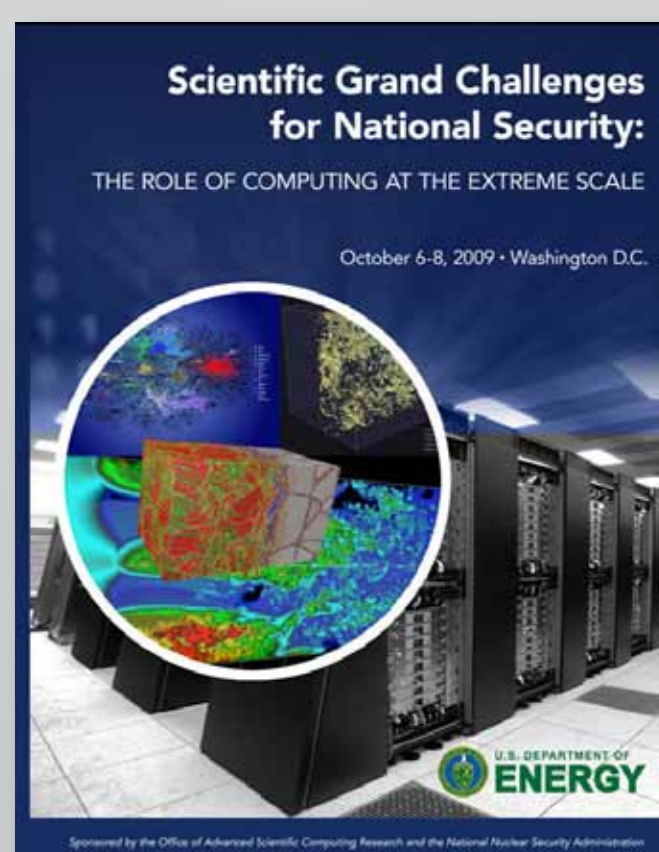
1997



2006



2007



2009



2010

Tim Germann, tcg@lanl.gov